DOCUMENT RESUME

ED 163 056 TM 008 108

AUTHOR Burton, Nancy W.

TITLE Performance Standards. Paper #11 in Occasional Paper Series.
INSTITUTION Western Michigan Univ., Kalamazoo. School of Education.
SPONS AGENCY Carnegie Corp. of New York, N.Y.
PUB DATE Dec 77
NOTE 25p.
AVAILABLE FROM Mary Anne Bunda, The Evaluation Center, College of Education, Western Michigan University, Kalamazoo, Michigan 49008 ($2.00)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Academic Standards; *Basic Skills; *Criterion Referenced Tests; Decision Making; *Educational Accountability; *Evaluation Criteria; *Performance Based Education; Performance Tests; Student Evaluation

ABSTRACT

Performance standards are useful tools when comparative evaluation or personal selection are not appropriate for decision making. The concept was developed in 1962 when Glaser introduced criterion referenced testing, to emphasize the importance of making test scores informative about learning processes rather than comparing them to other people's scores. It was hoped that criterion referenced testing, with performance standards, would provide a simple, inexpensive decision making tool. There are three methods of setting performance standards that might allow us to relate tests to learning procedures: standards based on theories, on expert consensus, and on minimum competencies. Learning hierarchies and other theories are too simplistic; expert standard setters lack the political thrust necessary to support decisive action. The minimum competency concept of performance standards is inadequate because no single skill, whether a prerequisiste for learning or for daily existence, is so essential that it can be defined as necessary for survival. No performance standard technology exists or is likely to appear, despite the growing amount of legislation depending on such a technology. Educational technologists should search for other methods to help decision makers and inform the lay public. Criterion referenced testing has not satisfied accountability needs; partly because performance standards are so arbitrary. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED163056

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Performance Standards

by

Nancy W. Burton

December 1977

EVALUATION CENTER

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

M. A. Burda

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

COLLEGE OF EDUCATION
WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

Paper #11
in
Occasional Paper Series

IM008 108

WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

OCCASIONAL PAPER SERIES

The Occasional Paper Series is published by the Evaluation Center, College of Education, Western Michigan University. It provides information about new developments in evaluation to various educational audiences.

For information about previous publications in the series, see the inside of the back cover of this document.

All correspondence in reference to the Occasional Paper Series should be addressed to:

Information Officer
Evaluation Center
College of Education
Western Michigan University
Kalamazoo, Michigan 49008

EVALUATION CENTER

Daniel L. Stufflebeam
Director

Mary Anne Bunda

James R. Sanders

ABOUT THE CENTER

The Evaluation Center is committed to advancing the science and practice of evaluation. To serve its purpose the Center has developed components in research, development, instruction, dissemination, and service. Over the past ten years the staff of the Center has had significant experience in these areas working with school districts, state departments of education, foundations, and federal agencies.

The *Research Component* is designed to develop and test evaluation theory. Particular thrusts are testing the effects of school-based evaluation offices on school practice, and determining the special evaluation skill and knowledge profiles required by various groups in education. The *Development Component* is designed to work on ways and means of implementing evaluation theory. Included are techniques for meeting special evaluation objectives, new data collection instruments, and special procedures for communicating evaluative findings and relating them to practical problems. The *Instructional Component* is designed to provide in-service training in evaluation to various local, state, and national groups. While the Center is not a credit-granting agency, it collaborates with instructional units throughout Western Michigan University for the development and implementation of pre-service evaluation curricula. Emphasis in this work is at the doctors and masters levels. The Center serves as an instructional laboratory wherein students majoring and minoring in evaluation can gain practical evaluation experience. The Center is also committed to the development of novel and needed evaluation training materials and strategies. The *Dissemination Component* is designed to communicate research findings, new ideas, and materials to the education profession. *The Service Component* is designed to assist agencies in meeting their evaluation needs through two primary services. These are the design and implementation of evaluation studies and the development and installation of agency evaluation systems.

The staff of the Center has been involved in conceptualizing the CIPP Evaluation Model, researching the Item Sampling technique, developing the Advocate Team Technique, developing over 100 standardized tests, designing and conducting many in-service evaluation programs in evaluation. Other accomplishments include authoring the PDK book, EDUCATIONAL EVALUATION AND DECISION MAKING, and numerous journal articles and papers, designing and conducting many evaluation projects, and installing evaluation systems in schools, government agencies, and research agencies. Overall, the Center staff has directed externally funded projects in evaluation totaling more than four million dollars.

The educational technologies of the past quarter century from teaching machines to minimal competency testing -- all share the general purpose of helping educators make better, or at least more uniform, decisions. The main support for these technologies has come from centralized agencies like the Office of Education, philanthropic foundations, publishers or state departments of education which wish to influence a system of education with extremely decentralized control. Since these central agencies cannot make local decisions, they instead appear to be trying to influence the parameters within which decisions are made.

The currently favored technique for shaping local decisions is criterion-referenced testing. Some criterion-referenced testers first find out how well students can read or write or play musical instruments and then tell the decision maker how well students <u>should</u> read, etc. The concept of performance standards has general appeal. All of us would like to know how many high school seniors are functionally literate or how many kindergartners are ready to read.

In this paper I will argue that no practical performance standards technology exists yet (despite the growing amount of legislation that depends on such a technology), and, furthermore, that the potential for such a technology is extremely limited. I conclude that educational technicians have an obligation to search for other methods to meet the needs that performance standards proposed to meet.

<u>The need for performance standards</u>

The demand for performance standards comes from two slightly different sets of needs. The first demand is to help educational decision makers; the second is to provide some comprehensible information to the lay public -- to parents, journalists, voters. The demands overlap to

the extent that laymen (school board members, legislators) are called upon to make educational decisions.

The public need is easily stated. The world is complex. Educational goals and options are expanding: the schools offer calculus, literature, sex education, vocational counseling, job training, breakfast. Educational information is exploding: per pupil expenditures are rising, tax bases are eroding, dropouts are increasing, college board scores are falling, average IQ's are rising, kids can't balance a checkbook, there's a surplus of teachers, learning grammar inhibits creativity, school grades don't correlate with income and so on. The public has a great desire to know some rather simple information: is my son or daughter getting a good education? Are the schools producing graduates qualified to meet society's needs for scholars and sanitation workers? Do I need to worry about the schools?

Performance standards offer one kind of answer to those general social questions.

The primary demand for standards, however, comes from educational decision makers. Their needs are more specific than the public's. The public makes very general decisions. Can I afford to vote against this school bond issue? Should I write to my congressmen about the spelling skills of high school graduates? Educators must not only decide what the problems are but also how to solve them. They need information that tells them which solution is best or at least which small set of solutions is likely to be best.

The major purpose of this paper is to consider how performance standards might help decision makers meet their needs. I treat the public need for information as secondary. If performance standards

could be developed to meet decision makers' needs they would also, as a by-product, meet the more general public need. If, as I argue, no such performance standards technology is likely to be developed, the general public need should be reconsidered. The concept of performance standards -- which has been aimed mainly at educational decisions -- may be a relatively poor way to meet the general need alone.

The development of the concept of performance standards

The most powerful distinction in contemporary educational testing is that -- originally posed by Robert Glaser -- between norm-referenced and criterion-referenced tests. This concept became widely known when Glaser published "Instructional Technology and the Measurement of Learning Outcomes" in the American Psychologist in 1963.

Glaser (1963) sought to emphasize the importance of making scores informative about behavior rather than merely about relative performance on poorly-specified and vaguely-known dimensions assumed to lie behind a test score:

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. The term "criterion," when used in this way, does not necessarily refer to final end-of-course behavior. Criterion levels can be established at any point in instruction where it is necessary to obtain information as to the adequacy of an individual's performance. The point is that the specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.

> Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which asses student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others. (pp. 519-520)

Glaser's original use of the word "criterion" is very close to the usual measurement usage:

> When we estimate a test's validity we must know which trait we wish the test to measure. This trait is called the criterion variable. We are interested in knowing how well the individuals' positions on the obtained distribution of scores correspond to the individuals' positions on the continuum which represents the criterion variable. (Magnusson, 1966, p. 124)

There was, however, a crucial difference in emphasis between Glaser's and the traditional measurement position. Validity (as the word implies) has primarily been an ethical requirement of tests, a prerequisite guarantee, rather than as an active component of the use and interpretation of tests. For example, the 1974 APA/AERA/NCME Standards for Psychological and Educational Tests refer to validity in this way:

> B5. Evidence of validity and reliability along with other relevant research data, should be presented in support of any claims being made. (p. 16, emphasis added.)

Glaser was, in essence, taking traditional validity out of the realms of research into the active arena of test use. He told the testing community that educators not only need to know that the scores on a test are positively correlated to important behaviors (the certification aspect), they need to know specifically that a given score implies a specific level of competence which leads them to make a certain decision.

Precisely how one makes this decision is one of the watershed issues in the development of thought on criterion-referenced testing. One must select a cut-off point on the test scale beyond which children may go on to the next activity. The cut-off point could be chosen using traditional selection techniques: namely, estimating probabilities of success in the next unit (based on the correlation between test and future promotion) and perhaps also considering such variables as student attitudes or costs. This approach to decision making is well developed in the traditional literature of measurement. It has a limited application simply because only those decisions which are general (college selections) or large scale (selection to flight training) could justify the complex apparatus of validation.

The criterion-referenced testing movement has become so important in large part because its proponents were able to recommend a much simpler decision procedure. At almost the same time that Glaser was calling for tests closely related to criterion behaviors, Robert Mager (1962) was propounding the idea of instructional objectives.

> If we can specify at least the minimum acceptable performance for each objective, we will have a performance standard against which to test our instructional programs: we will have means for determining whether our programs are successful in achieving our instructional intent. (p. 44)

It was inevitable that these two ideas would be combined. Given a <u>criterion-related</u> test (in the older "validity" sense of the term) with a <u>criterion</u> score, data-based decisions were suddenly reasonably simple and cheap. Instructional technology became much more practical for the individual classroom. The problems of program evaluation, assessment, and accountability -- all closely-related technology-based contemporary movements -- were greatly simplified.

It has become well accepted that the criterion is the cut-off score, that its main virtue is in making some tough distinctions.

> A criterion-referenced test should be designed to divide students into just two groups -- those who have mastered the material and those who have not. (Greenbaum, 1977, p. 82)

Criterion-referenced testing ideas have come to be closely associated with the idea of accountability which began to grow rapidly at about the same time. The growth of the accountability movement placed a great deal of pressure on the testing industry. Conventional standardized, norm-referenced tests were seen as inadequate for the new purposes. As a result, the new term "criterion-referenced" became something of an umbrella, covering the attempts to correct all the sins of past testing practice. However, it seems clear that criterion-referenced tests are as susceptible as norm-referenced tests to excessive concentration on specific factual knowledge and might easily employ biased items or have unclear directions or formats.

There are several unique attributes of norm-referenced tests that restricted their usefulness for accountability/assessment. Two of the most important occur because these tests:

- Measure rather general achievements not associated with any particular curriculum. As a result, they tend to be closely related to ability measures and not very sensitive to changes (either improvements or mistakes) in curriculum.

- Include items which have been selected mainly for their effectiveness in spreading kids along a continuum high to low achievement. Such items may not help decision makers decide what skills kids do or don't have.

Criterion-referenced tests, however, introduce a new problem. With a norm-referenced test, one can always interpret a score in light of the norms group: Johnny does better than 85% of his class, or better than the typical eighth grader. In contrast, it is not so easy to

decide whether Johnny's ability to solve 80% of all 2-digit addition problems without carrying is good or bad. How good does he need to be at addition before he can learn to multiply? Or before he can keep his checkbook in reasonable order? The answer to these interpretation problems was, obviously, performance standards.

Nearly fifteen years have passed since the idea of criterion-referenced testing was adopted as a most promising way to provide accountability in education. The original hope, that criterion-referenced tests (with performance standards) would provide a simple or inexpensive tool for decision makers, has not been realized. Glass (1977) has done an extensive review of techniques for setting performance standards, and concludes that all are dangerously arbitrary. Shepard (1976) concludes that present procedures for setting standards all reduce to a poor form of norm-referencing. Decisions are made relative to the standards-setter's personal experience of childrens' performances, rather than a carefully-selected representative sample of performances.

The time has come to look at the assumptions and purposes of criterion-referenced tests to determine whether they are ever likely to become possible.

Some distinctions about criterion-referenced tests

I have already mentioned that "criterion-referenced" has become a rather vague umbrella term. However, nearly all discussions of criterion-referenced tests assume that they are part of a decision procedure that has, in Robert Glaser's words, "action consequences." This definition excludes certain certification uses of tests. For example, gold stars for learning scales are excluded unless the teacher

won't let students learn tunes until they have all their gold stars; high school diploma competency requirements are excluded unless employers make hiring decisions based on the possession of a diploma.

The kind of decision appropriate to criterion-referenced technology is a medium-size, midrange decision. The day-to-day decisions of education are too numerous and to short-lived to justify the entire apparatus of empirical investigation. The broadest educational decisions are based much more on widespread public beliefs about education (true equality of educational opportunity requires compensation for poor social or economic background) and political realities (affluent school districts must not be completely cut off from Title I, ESEA monies) than they are on empirical information.*

There are many mid-size decisions that can appropriately be based on performance data. For an individual child, the teacher might decide to skip three lessons or repeat one, to try a different instructional strategy or send him or her to another class. A principal might decide to reassign teachers or hire aides or create or disband a low-ability track or to require all classes to spend an hour a day on reading.

These mid-level decisions have traditionally been the most successful arena for educational technologies. Various comparative techniques exist for assisting in those decisions that can be characterized as limited-selection decisions: a school district must decide on which set of curriculum materials to purchase or a college must decide on which one thousand freshmen to accept. A large literature on compar-

*See N. W. Burton, "Assessment and Social Policy," (1977) for a discussion of how empirical information can influence general social policies. Briefly, I argue that empirical information helps form a view of the world or paradigm which both dictates the problems that will be attended to and the kinds of solution that will be proposed.

ative evaluation and on personnel selection provide methods for making such decisions. The literature of performance standards has sometimes included these older, well-established comparative techniques. (See Glass, 1977). In this paper, I am concerned with that range of educational decisions for which comparative techniques are not appropriate but which performance standards attempt to address. Generally, that class of decisions includes all non-limited selections: How many remedial or advanced-placement classes should the school offer? How many students should pass sophomore English? (See Cronbach and Gleser, 1965).

Note that there are two steps in non-limited selection. Selection is to be followed by treatment. In the limited case, there is one treatment (say, college) and one need only decide who is to get that treatment -- it is a one-step decision. In the non-limited case, there are multiple treatments and the problem is to assign children to the correct one. Thus one is not simply identifying students who need help -- one is identifying students who need a particular kind of help. In the literature of performance standards the problem has usually been stated as the single problem of identifying those who need help -- as if the type of help was an entirely different question. I am arguing that standards should diagnose, not simply identify.

My argument stands on two legs. First is the logic of the situation. To pursue the medical analogy, one does not diagnose disease by looking for disease in general. After disquieting symptoms are discovered, there is the attempt to match the patterns of symptoms to the known patterns of specific diseases. The second leg of my argument is utilitarian. There is no need to diagnose unless diagnosis usually leads to treatment. There is no use for performance

standards that don't help us solve problems.

Tests are only one part of the decision making model that leads to treatment of problems. However, most of the literature has concentrated exclusively on techniques for providing desirable measures, assuming that decision makers could make better decisions if appropriate tests could only be devised. I believe that the testing community have ignored an extremely important part of the decision making situation in its concentration on the technicalities of testing. Namely, testers have ignored the connection between performance standards and action alternatives. Perhaps the reason is, as Arthur Wise (1977) remarks,

> educational policy is designed to alter the practice of education without an understanding of how education actually occurs. ...(P)olicy makers may believe that it is sufficient to cause something to occur by legislating that it should occur. (p. 15)

However, standards are much more closely tied to action alternatives than they are to tests. The fact, for instance, that a 70% individual success rate on addition problems is a sufficient background for success in multiplication doesn't add to your information about children's addition skills, but it does help you to decide what to <u>do</u> about any given level of success on an addition test.

By clarifying this association with actions one can better understand the leading characteristic of criterion-referenced tests, the fact that criteria are expressed in terms of <u>tasks</u>. This is because of the relation of criteria to learning processes, The criteria that make sense are not comparisons with other people's performance (that is, they are not norm-referenced) because criteria are related to <u>actions</u> (that is, to learning processes).

Of all the methods of setting performance standards for criterion-referenced tests, there are three that might possibly allow us to relate

tests to actions. They are:

(1) Standards based on theories.

(2) Standards based on expert consensus.

(3) Standards based on practical necessities.

Standards based on theories

The original conception of criterion-referenced testing was closely related to learning hierarchies.

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by the behaviors he displays during testing. (Glaser, 1963, p. 519)

What is this continuum of achievement? The clearest interpretation is of a learning hierarchy: where each step or set of steps is the necessary and sufficient condition of the following steps. If test scores could be mapped onto this hierarchy of behaviors, then each final score would imply that the respondent can do the corresponding task and all tasks below it in the hierarchy.

Unfortunately, any counterexample is enough to upset the purity of the hierarchy. Only the most trivial examples of hierarchies have ever been advanced. In fact, most models for learning do not posit such a simple, linear, step-by-step method of learning. "It seems probable that only very narrowly defined abilities can be used in establishing a hierarchical relationship" (White, 1973, p. 371).

One need not abandon achievement continua simply because one doubts that they reflect the true structure of learning. Whatever the internal processes, it could still be argued that learning proceeds linearly because people can learn only one thing at a time. Are there any regularities in the order of learning? Is it not true that adding

is usually taught before multiplying? A de facto, as opposed to a theoretical, scale could be created for any learning area that is usually taught in a particular sequence. De facto scales could also be constructed for any area that is usually mastered in a particular sequence. Most people, for example, learn to read easy passages before they do difficult ones; master walking before they dance; understand concrete operations before abstract. However, there is a serious measurement problem with either theoretical or de facto scales. The problem is that the difficulty of an item does not depend only on the complexity of the concept being measured. Items can be written about a passage from Kant that 90% of the people can pass and items can be written about an Ann Landers answer that few could pass. This is not a trivial technicality. The profound truth in such claims as Bruner's (1960) -- that he can teach any concept at any age -- is that concepts have many levels of meaning. Einstein could explain the general theory of relativity to a bright high school student. As a result, scaling by simply relating test items to a continuum of concepts is impossible. Scale scores would not correspond to the continuum.

Furthermore, even supposing that an empirical scale could be constructed, results on that scale would be an ambiguous basis for decisions. Suppose the children in District 1 did not attain the usual level by third grade. The lack of attainment could mean a defect in instruction; it could also mean simply that District 1 did not follow the usual sequence of instruction. (There is no necessity involved in de facto hierarchy.)

There is a great deal of work going on now with non-hierarchical learning theories. At present, none of these theories are broad enough to be considered as general decision making tools. However, such

theoretical areas as psycholinguistics could be examined now for appropriateness in setting performance standards in reading. Since the areas of application of any such theory would likely be restricted, this effort would only be useful, at this point, as a pretest of a performance standards decision model.

Standards based on expert consensus

Lacking a theory, experts may still set standards based on their experience. Shepard (1976) has presented an excellent discussion of the problems of basing standards on experience.

> But what do we mean by "experience" but imperfect norms? Each expert is likely to base his opinion on his own sixth-grade experience, or that of children he has taught or that of his own sons and daughters. For standards to be realistic or attainable, experts are likely to reflect on what current sixth-graders are able to do; by relying on experience, their standards will be normative but will lack the representativeness of more systematic sampling. (p.6)

As in the case of empirical scales, these judgmental standards based on experience lack several important features. Since they are based on comparisons with reality, both the justification and the consequences of the standard must be established by some external evidence, since they do not follow logically from a theory.

Shepard (1976) advocates giving judges all the evidence they can use.

> Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms. ... Of course, the norms are not automatically the standards. Experts still have to decide what "ought" to be; but they can establish more reasonable expectations if they know what current performance is than if they deliberate in a vacuum. (p. 11)

Given the empirical evidence they lack, expert standard-setters have significant advantages over the scaling method discussed above. The experts' authority helps to justify the standard. And, lacking

theoretical or empirical information about causes, the judges' experience can still provide some speculations about appropriate actions to follow from passing or failing the standard.

The standard the experts set will essentially be the empirical "norm" plus something for education aspirations, plus or minus something more, depending on the severity of positive or negative errors. One could hardly convince himself that such a standard can define the difference between masters and nonmasters. Such standards hardly seem the basis for hard-headed educational decisions. And, in fact, everywhere that such standards have been imposed, the consequences of failure have been so softened that it is impossible to evaluate their effect.

So far, I have discussed several logical and several empirical ways of setting standards. I dismissed the idea of learning hierarchies as too simplistic and other learning theories, at present, as too limited to offer much practical guidance. De facto scales that would look like hierarchical scales but lack the objectionable theoretical trappings were briefly discussed; I could not determine how much scales could by themselves perform the function of standards since they had neither intrinsic nor extrinsic justifications and they did not imply action consequences.

I then turned to experts' judgments. However, once again, I could not see how experts' judgments could fill the need for performance standards. Since performance standards exist in a political context -- that is, they exist to help decision makers act on policy matters -- they require the most careful justification if they are to have any discernable consequences. The simple authority of expert judgment has not so far been sufficient.

Standards based on practical necessities

In the early seventies, a different concept of performance standards began to appear. Legislators and educators began to talk of "functional skills" or "minimal competence." This development was based on the idea that once one can identify the skills needed in everyday life, the practical value of the skill becomes its justification as a standard. As Naomi Rosh White (1973) saw it, this new idea is not entirely different from the earlier ideas of criterion-referenced testing, but it is an important shift in emphasis.

> Broadly speaking, minimal skills can be said to have two meanings. These meanings are not mutually exclusive, but represent differences in emphasis.
>
> "Minimal" as "Prerequisite for Learning"
>
> This usage is predicted on an ordinal or hierarchical notion of learning. That is, the student, in order to come to know certain things, must possess basic skills which will facilitate acquisition of further knowledge. These skills are necessary antecedents to cognitive activity, antecedents both temporally and logically.
>
> It would seem to follow, then, that the appropriate perspective for exploring this interpretation of minimal skills would be cognitive and learning theories. These theories might provide evidence for what basic skills are, how one acquires them, and how one might test for them.
>
> "Minimal" as "Necessary for Personal Efficacy"
>
> This usage is predicated on the view that everyday existence requires of each person skills necessary for survival. That is, a person, in order to function effectively in the social environment, must possess certain skills. It further implies that one is fully a person only if one possesses these skills. The skills are necessary and sufficient for -- personhood -- self-actualization ... whatever terms one cares to choose. This stands in marked contrast to the usage described above, where the skills are necessary for learning, but not sufficient. (p. 158)

Minimal competency language seems best suited to simple yes/no situations: either you can swim or you sink. It can also be applied to situations where one can imagine a continuum of skill. With a continuum, one needs to posit some threshold level of skill or critical mass of facts below which one could not be competent as a citizen, a carpenter, a medical doctor -- at any rate, competent in some real-life role.

Can this threshold be arbitrary? For example, Standard Oil of California will hire no typist who scores less than 50 words per minute. Why can't other performance standards be as arbitrary? Such a standard could not be justified, for example, in a situation where the potential typists had any recourse to question the validity of the standard or whenever the demand for typists exceeded the supply.

Any practical standard must be seen as real. But then the standard-setters are plagued by exceptions. Shepard (1976) considers this problem.

> They might, for example, interview plumbers and shop clerks to try and locate which competencies are minimal, which are held by all employed adults. Such searches will be informative and may be helpful to judges in the same way normative data are useful, but the searchers will not turn up any universals. We might as well consider right now what we will do with some very successful plumbers who cannot read or street sweepers in San Francisco who make more money than university professors but cannot handle simple fractions. The search for absolutes leads to absurd reductionism. Perhaps we should study the mentally retarded. What skills are absolutely essential to be able to ride the bus to and from a sheltered workshop? Suppose we thus identify some basic skills that are completely rudimentary. How would these lowest-common-denominator standards serve as meaningful criteria for prospective carpenters and TV repairmen?
>
> Standard setters ought to begin their task recognizing that counter examples will exist. If reading comprehension is deemed important, the standard ought to be set despite the existence of successful businessmen who cannot pass the test. Lowering the standards until everyone can pass completely defeats their purpose. (pp. 6-7)

Shepard recommends that standards be set despite their ambiguity. I hold that this kind of ambiguity will also defeat the purpose. If such standards as diplomas are used to make employment decisions, we will be unable to withhold Steve's high school diploma for something that Steve's father's boss didn't need to be successful in life.

Besides, these counterexamples are not merely irritating specks on an otherwise perfect surface. There is something fundamentally wrong with the idea of using specific achievement measures when thinking about success or failure in any real venture. My success in life cannot be completely determined by how many reading or science or music chips I have. And it won't help to count my attitudes, my beauties, or my dollars. Success in any venture may come from anywhere in my repertoire of skills. I can read well enough or I can talk well enough or I can learn to avoid the problem. The fact that I can or cannot read at some level will never, by itself, completely determine my success. This is the meaning of the following aphorism attributed to John Tukey by Gene Glass (1977).

> "Life is like a Double-crostic: we can do far more than we know." When one first reads the definitions of the words in the Double-crostic, he discovers that he knows only a half-dozen or so among fifty or sixty. But eventually, through the complex and interlocking system of semantic and linguistic clues of the puzzle, all of the words and the quotation are identified. (p. 30)

This insight fits well with current learning theories. Organic metaphors are being explored rather than the early stair-step models. Growth in learning is related to the complexity and generality of integration among concepts. Success in any venture depends on the entire conceptual network.

Given this outlook on learning, the issue of performance standards loses potency. No measure of a single skill can ever be mapped onto a

nontrivial vision of real success because any problem can be solved in more than one way. One can determine whether the respondent has the skills necessary to solve the problem this way, but one lacks the justification for imposing successful performance, this way, as a standard.

The argument applies to more than the practical necessities of daily life. It applies just as well to algebra problems and sculptures. It applies to any reasonably complex task in which one could succeed or fail. I believe this argument is fatal to any method of setting performance standards, including any conceivable learning theory. The argument is that performance standards are not sensible for any problem that has more than a small, definable set of possible solutions. But such restricted problems are almost by definition trivial. Performance standards simply cannot help us decide whether Johnny or PS 19 or Colorado has enough reading skill, because there is no sensible answer to the question, "Enough reading skill for what?", beyond the trivial level of "Enough reading skill to answer test question 36b correctly."

The discussion of minimal skills reflects back on the entire standards question. Suppose, for example, I were teaching conservation of volume using Piagetian theory. I might try to construct performance standards to monitor progress toward the concept of conservation. I might even be able to say "Unless the student can answer 7 out of 10 questions on X, he or she will never get to the final concept in this course." But this becomes a very limited statement. The student may learn conservation of volume some other way or he may adjust successfully to never having the concept. At any rate, the performance standard will only be applicable within the context of my curriculum unit. Even within the classroom, the reason for insisting that

students learn the concept my way is not for the students' good or the good of society -- it's for my convenience as a teacher. Within my classroom, performance standards are part of my instructional strategy. They tell me little about the quality of my students and they tell you nothing about the quality of my teaching: they are just signposts guiding me through my chosen instructional strategy.

This, I take it, is close to Glaser's original purpose for "criterion-referenced testing." Since criterion-referenced tests were immediately appropriated by the accountability movement, this original purpose may never have been given an adequate test. My argument in this paper has simply been that criterion-referenced testing has been a disappointment for purposes of accountability.

Conclusion

I have attempted to show that the major purpose for performance standards is to help educators make decisions. Specifically, performance standards are of use for making mid-size decisions in those situations where the techniques of comparative evaluation or personnel selection are not appropriate. To be useful for these decisions, the performance standard should imply some action or set of actions to follow from acceptable and unacceptable performance and the standard should invoke some authority to support the decision.

I presented three procedures for setting performance standards that were likely to meet these criteria. Theoretical procedures were rejected because learning hierarchies have never been established and other theories seem at present too limited. Expert judgments were rejected because at present they do not appear to be politically compelling enough to support decisive action. Finally, "minimal

competency" techniques were rejected because real-life successes have many potential causes. No single skill is so essential that it can be defined as necessary for survival.

The basic problem in using criterion-referenced testing for accountability is in generalizing from the instructional setting:

Johnny must do A, B, & C to get through the course my way

to:

Johnny must do A, B, & C to get through life.

There may be some "basics" that are necessary to survival, but they are on the order of life, liberty, food, and shelter. There is another set of basics that is quite inconvenient to do without -- such as hearing and, to a lesser extent, reading. However, even these most basic skills can do no more than affect the probability that one will have a comfortable life. It is true that having skills is good and the more skills the better. The more skills one has the greater one's chances of success by any number of measures. However, it does not therefore follow that there are some skills so necessary that they can be called "standards" and demanded of everyone.

References

American Psychological Association. Standards for educational and psychological tests. (Rev. ed.) Washington, DC: APA, 1974.

Bruner, J.S. The process of education. New York: Random House, 1960.

Burton, N.W. Assessment and social policy. Unpublished draft, 1977. (Available from the author, Education Commission of the States, 1860 Lincoln, Denver, CO, 80295).

Cronbach, L.J. & Gleser, G.C. Psychological tests and personnel decisions. (2nd ed.) Urbana, IL: University of Illinois Press, 1965.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glass, G.V. Standards and criteria (The Evaluation Center's Occasional Paper No. 10). Kalamazoo, MI: The Evaluation Center, Western Michigan University, 1977.

Greenbaum, W. Measuring educational progress: a study of the national assessment. New York: McGraw-Hill, 1977.

Mager, R.F. Preparing instructional objective. Palo Alto, CA: Feardon Publishers, 1962.

Magnusson, D. Test theory. Reading, MA: Addison-Wesley, 1966.

Millman, J. Domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Shepard, L.A. Setting standards and living with them. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1976.

White, N.R. What are schools for? In M. Clasby, M. Webster, & N. White. Laws, tests and schooling. Syracuse, NY: Educational Policy Research Center, Oct. 1973.

White, R.T. Research into learning hierarchies. Review of Educational Research, 1973, 43, 361-375.

Wise, A.E. Why educational policies fail: the hyperrationalization hypothesis. Journal of curriculum studies, May 1977, in press.

Additional copies of papers in this series are available upon request for a fee of $2.00. Previous papers in this series are listed below. Please direct your inquiries to

Mary Anne Bunda, Series Editor
The Evaluation Center
College of Education
Western Michigan University
Kalamazoo, Michigan 49008

1. A Response to the Michigan Education Department's Defense of Their Accountability System by Daniel L. Stufflebeam.
2. An Administrator's Perspective of Evaluation by Jack P. Taylor.
3. Meta-Evaluation by Daniel L. Stufflebeam.
4. Evaluation Bias and Its Control by Michael Scriven.
5. Program Evaluation, Particularly Responsive Evaluation by Robert E. Stake.
6. A Basis for Determining the Adequacy of Evaluation Designs by James R. Sanders and Dean H. Nafziger.
7. Needed: Instruments as Good as Our Eyes by Henry M. Brickell.
8. Assessing the Impact of Planned Social Change by Donald T. Campbell.
9. A Checklist for Evaluating Large-Scale Assessment Programs by Lorrie A. Shepard.
10. Standards and Criteria by Gene V Glass.
11. Performance Standards by Nancy W. Burton.